Lätti *et al. J Cheminform (2016) 8:45*
DOI 10.1186/s13321-016-0158-y

Journal of Cheminformatics

SOFTWARE

Open Access

Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization

Sakari Lätti, Sanna Niinivehmas and Olli T. Pentikäinen*

Abstract:
Receiver operating characteristics (ROC) curve with the calculation of area under curve (AUC) is a useful tool to evaluate the performance of biomedical and chemoinformatics data. For example, in virtual drug screening ROC curves are very often used to visualize the efficiency of the used application to separate active ligands from inactive molecules. Unfortunately, most of the available tools for ROC analysis are implemented into commercially available software packages, or are plugins in statistical software, which are not always the easiest to use. Here, we present Rocker, a simple ROC curve visualization tool that can be used for the generation of publication quality images. Rocker also includes an automatic calculation of the AUC for the ROC curve and Boltzmann-enhanced discrimination of ROC (BEDROC). Furthermore, in virtual screening campaigns it is often important to understand the early enrichment of active ligand identification, for this Rocker offers automated calculation routine. To enable further development of Rocker, it is freely available (MIT-GPL license) for use and modifications from our web-site (http://www.jyu.fi/rocker).

Background

In early stages of drug discovery, virtual screening (VS) offers an attractive way to identify hit molecules for the target protein. Although there are a wide variety of tools to perform VS, it is necessary to validate their efficiency in separation of active ligands from inactive molecules. One issue that has helped validation significantly is the appearance of databases of ligand binding data, e.g. ChEMBL [1], and molecule collections, where not only active ligands but also decoy molecule sets are available, e.g. DUD [2], DUD-e [3], and DEKOIS [4, 5]. The other important issue in VS efficiency is the numerical and visual illustration of how well the VS method works. For this, two issues are typically calculated: (1) area under curve (AUC) for the receiver operation characteristics (ROC), and (2) early enrichment, e.g. upon the top 1 %. There are many possibilities to avoid the bias in the ROC AUC analysis [6, 7]. The ROC AUC value itself does not directly give detailed information about the early enrichment, but the visualization of it does. Especially, plotting ROC as a semi-logarithmic curve improves the readability a lot. Also weighting each active based on the size of the lead series to which it belongs [6] or incorporating the notion of early recognition into the ROC metric formalism [7] can give useful information about the enrichment of the active molecules. When ROC AUC value is reported with early enrichment, already the two numbers give a good idea for the quality of the used method to separate true positives from false positives.

For the ROC AUC visualization there are many tools [8], e.g. pROC [9], ROCR [10], Pcvsuite [11] that work on top of widely used R-package, and some of them contain sophisticated ROC comparisons for the analysis of medical data. Furthermore, there are web-based tools, such as jrocfit (http://rad.jhmi.edu), and standalone tools like MedCalc [12]. However, as all of these tools have been developed for calculation and comparison of medical data, they do not continue handy tools for VS efficiency analysis. Furthermore, the VS efficiency data is used in the comparison of different VS strategies and tools, and as we noticed in our previous study

*Correspondence: olli.t.pentikainen@jyu.fi
Computational Bioscience Laboratory, Department of Biological and Environmental Science and Nanoscience Center, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

[13], authors have different opinions about the methods and types of calculations that should be employed with VS analysis. Motivated from this, we introduce a very user-friendly tool called Rocker dedicated for the VS analysis. Rocker calculates the ROC AUC-values, BEDROC-values [6, 7], draws the curves either as semi-logarithmic or non-logarithmic scale, and calculates the enrichment at the given percentage with two commonly used ways.

Implementation

Rocker is written with Python, and requires in addition to that, the Python-matplotlib library, which is typically available through Linux package management tools, e.g. yum in Red Hat and Fedora distributions. The ROC and AUC are calculated using algorithms described by Fawcett [14]. Fawcett has described the algorithms in a clear way utilizing pseudocode. For the conservative estimate of the standard error for the AUC there are several solutions available, from which the commonly used method developed by Hanley and McNeil [15] was implemented into Rocker. Hanley's nonparametric approach has the advantage of being simple to calculate, and the corresponding accuracy indexes are obtainable even for small sample sizes [16]. Furthermore, the BEDROC-values with varied alpha can be calculated in order to calculate the ROC with weighted early enrichment [7].

Rocker can calculate the enrichment factors in two commonly used ways, in order to make it easier for the user to compare own results with the published ones: (1) for the top X % of the results, (EFX; Eq. 1), and (2) for the top results until X % of the decoy molecules have been found (EFXDEC; Eq. 2).

$$EFX = \frac{\frac{\mathrm{Ligs}_{X\%}}{\mathrm{Mols}_{X\%}}}{\frac{\mathrm{Ligs}_{all}}{\mathrm{Mols}_{all}}} \quad (1)$$

$$EFXdec = \frac{LigsX\%dec}{\mathrm{Ligs}_{all}} \times 100 \quad (2)$$

In Eq. (1) $\mathrm{Ligs}_{X\%}$, $\mathrm{Mols}_{X\%}$, Ligsall and Molsall are the number of the ligands in the top X % of the screened compounds, the number of the molecules in the top X % of the screened compounds, the total number of the screened ligands, and the total number of the screened molecules, respectively. In Eq. (2) $\mathrm{Ligs}_{X\%dec}$ is the number of the ligands when X % of the decoy molecules have been found and, again, Ligsall is the total number of the screened ligands.

There are some command line options available in Rocker to control the quality and properties of the output figure and to calculate the enrichment factor. The true and false positives can be separated in two ways: (1) true positives have some difference in their names, which is possible to indicate with regular expression, in contrast to false positives; (2) the names of the true positives can be given as list in a separate file. The figure itself can be manipulated in many ways: (1) resolution and size of the image can be adjusted; (2) labels, font and font size, including the location of legend can be modified; (3) X-axis can be drawn with linear or logarithmic scale; (4) axis thickness and tick size can be adjusted; (5) colors, thickness, and line styles (e.g. solid, dashed) of the plotted curves can be changed; (6) "random-selection curve" can be included or excluded. And finally, the enrichment factors can be calculated in two different ways (see above). The AUC is printed if sufficient data is given. If you do not have graphical output available, you can still calculate the AUC-values and enrichment factors by preventing the drawing of ROC.

Results and discussion

Rocker can be downloaded from http://www.jyu.fi/rocker for linux (rpm), windows, and mac os. Furthermore, Rocker can also be used via simplified web-interface (available at http://www.jyu.fi/rocker) where user can download the text-file that consists the name-field (1st column) and numerical data that describes the activity/fitness/score (column number for this data can be specified). In current web-interface version the names of true positives (or active compounds) should differ from those of false positives (or decoy molecules). Output figure can be drawn either with linear or logarithmic X-axis, ROC can be drawn either with solid or dashed line with option for color selection. Resolution of the figure can be specified. Furthermore, calculation of BEDROC, EF, and EFdec can be performed with wished values.

To visualize the performance of Rocker, here are six example commands, and the figures (Fig. 1) they produce from an example input files (found from Rocker homepage):

(A) rocker dude.txt -an CHEMBL -c 5 -s 5 5 -p Fig1A.png
This is simple example, where the names of all active ligands begin with CHEMBL (-an; Note that the names of inactives cannot begin with CHEMBL then). The 5th column has the score/fitness-value that is compared (-c; molecule names in 1st column). The produced image has is sized (5*5) inches (-s). Finally, the prepared figure is called Fig. 1a (-p)

(B) rocker dude.txt -an CHEMBL -c 5 -s 5 5 -lp 0.001 -p Fig1B.png
Similar as in (A) but the X-axis is drawn in logarithmic scale, beginning from 0.001 (−lp).

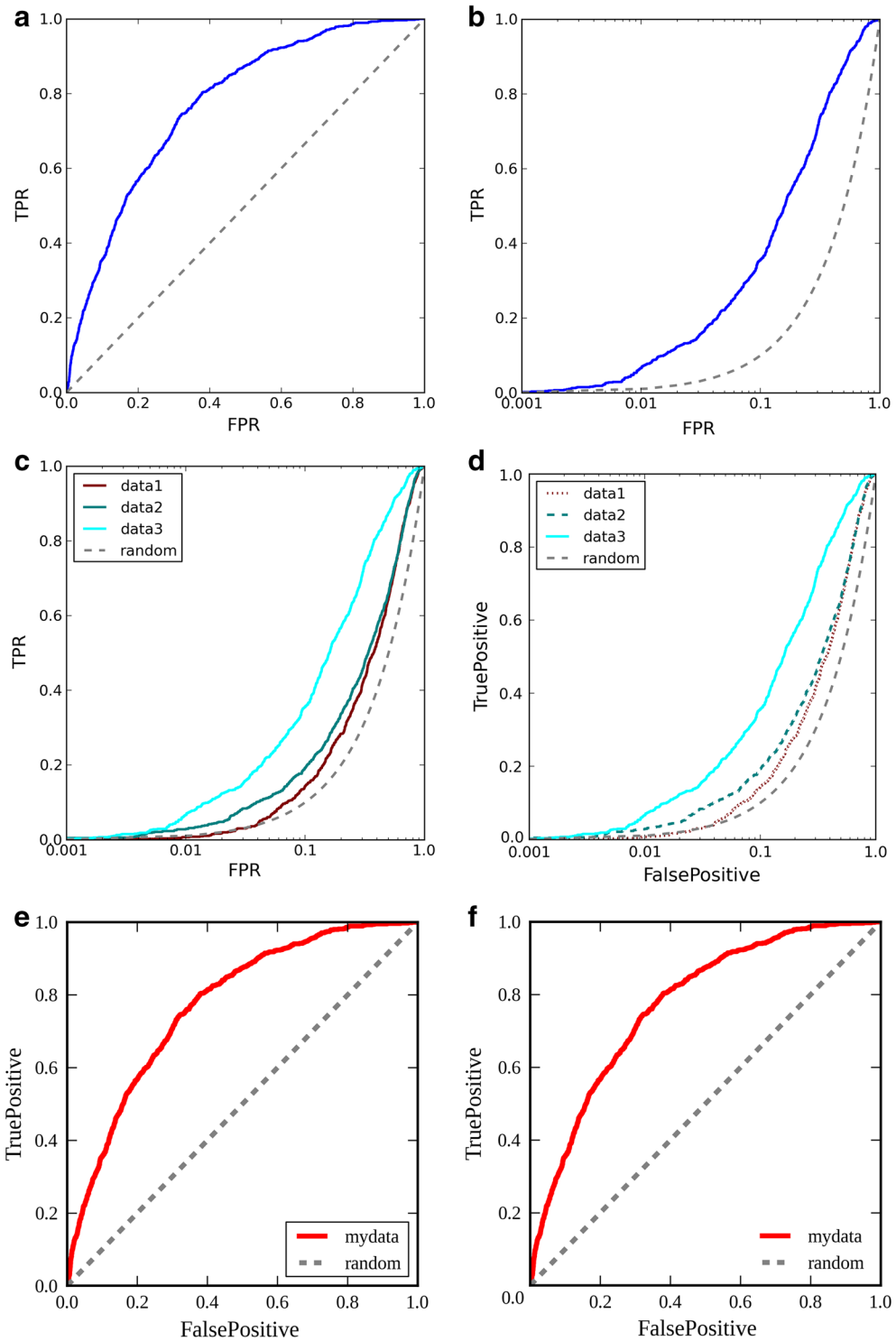

Fig. 1 Six example ROCs with different command line options

(C) rocker 2.txt 3.txt 5.txt -an CHEMBL -s 5 5 -li data1 data2 data3 random -l 0 -lp 0.001 -cl maroon teal cyan -p Fig1C.png

In this example, three curves are drawn from three different files (2.txt, 3.txt, and 5.txt). Legends for each curve are written (–li) and the position of leg-

end-box is indicated (-l). The colors of curves are given (-cl).

(D) rocker 2.txt 3.txt 5.txt -an CHEMBL -s 5 5 -li data1 data2 data3 random -l 0 -lp 0.001 -cl maroon teal cyan -st dotted dashed solid -la FalsePositive TruePositive -las 15 -ts 15 -p Fig1D.png
Otherwise as (C) but the line styles for curves are changed (-st), axis labels are defined (-la), and their font size is set (-las), label size for tick numbers is defined (-ts)

(E) rocker dude.txt -an CHEMBL -s 5 5 -c 5 -li mydata random -l 4 -les 15 -cl red -lw 4 -la FalsePositive TruePositive -las 15 -ts 15 -aw 2 -f "Liberation Serif" -p Fig1E.png
Here, the new elements are changed font size for legend (-les), linewidth of the ROC curves (-lw), width of axes (-aw), and defined font (-f).

(F) rocker dude.txt -an CHEMBL -s 5 5 -c 5 -li mydata random -l 4 -les 15 -cl red -lw 4 -la FalsePositive TruePositive -las 15 -ts 15 -aw 2 -f "Liberation Serif" -no -a '\-0.08,\-0.05,0.0' -as 15 -kw "legend:{frameon:False}" -EFd 1 -EF 1 -BR 20 -p Fig1F.png
Here, the origo is not drawn, i.e. values 0.0 for X- and Y-axis (-no), but one 0.0 is written to position coordinate position -0.08,-0.05 (-a), which font size is set to 15 (-as), the box for legend is not drawn (-kw). Furthermore, the enrichment factors (both types) are calculated at top-1% (-EFd, -EF), bedrock is calculated with alpha-value of 20 (-BR).

As an example, the output from command (F) (as well as output from web-interface), looks like this:

```
Loaded 860 actives from dude.txt starting with ['CHEMBL']
Loaded 28384 average scores from dude.txt
AUC = 0.773287667761 + −0.00949080870401
Plotting ROC Curve...
EF_1.0 = 5.22961021946
EF_1.0%decs = 4.76744186047
BEDROC_20.0 = 0.245830895735
```

Conclusions

As is, Rocker offers a highly useful, easy-to-use tool for ROC analysis in VS, including calculations of AUCs and early enrichments. Although authors sincerely hope that the future developments are made available for the other users as well, that is not required by the license.

Availability and requirements

Project name: Rocker.
Project home page: http://www.jyu.fi/rocker.
Operating system: Platform independent.
Programming language: Python.
Other requirements: Python-matplotlib.
License: MIT-GPL.
Any restrictions to use by non-academics: none.

Abbreviations

AUC: area under curve; BEDROC: Boltzmann-enhanced discrimination of receiver operating characteristics; ROC: receiver operating characteristics; VS: virtual screening.

Authors' contributions

SL wrote the code, SN and OTP tested the code and wrote the manuscript. All authors contributed into design of the study. All authors read and approved the final manuscript.

Acknowledgements

CSC, The Finnish IT Center for Science is acknowledged for computational resources (OTP; Project Nos. jyy2516 and jyy2585).

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data and materials are available at the project home page http://www.jyu.fi/rocker.

Received: 31 March 2016 Accepted: 1 September 2016
Published online: 07 September 2016

References

1. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(Database issue):D1100–D1107
2. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801
3. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55(14):6582–6594
4. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM (2013) Evaluation and optimization of virtual screening workflows with DEKOIS 2.0—a public library of challenging docking benchmark sets. J Chem Inf Model 53(6):1447–1462
5. Vogel SM, Bauer MR, Boeckler FM (2011) DEKOIS: demanding evaluation kits for objective in silico screening—a versatile tool for benchmarking docking programs and scoring functions. J Chem Inf Model 51(10):2650–2665
6. Clark RD, Webster-Clark DJ (2008) Managing bias in ROC curves. J Comput Aided Mol Des 22(3–4):141–146
7. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47(2):488–508
8. Stephan C, Wesseling S, Schink T, Jung K (2003) Comparison of eight computer programs for receiver-operating characteristic analysis. Clin Chem 49(3):433–439
9. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC et al (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 12:77
10. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940–3941

11. Pepe M, Longton G, Janes H (2009) Estimation and comparison of receiver operating characteristic curves. Stata J 9(1):1
12. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH (1995) MedCalc: a new computer program for medical statistics. Comput Methods Prog Biomed 48(3):257–262
13. Niinivehmas SP, Salokas K, Latti S, Raunio H, Pentikainen OT (2015) Ultra-fast protein structure-based virtual screening with Panther. J Comput Aided Mol Des 29(10):989–1006
14. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874
15. Hanley JA, Mcneil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) Curve. Radiology 143(1):29–36
16. Hajian-Tilaki KO, Hanley JA (2002) Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol 9(11):1278–1285